Comparación y evaluación de modelos desde la perspectiva del DM.

<u>Bibliografía:</u>

- Berthold, Michael; Hand, David. Intelligent Data Analysis.
- Hastie, T.; Tibshirani, R. y Friedman, J. The Elements of Statistical Learning.
- Giudici. Data Mining Model Comparison. Cap 32 de DM&KD Handbook. Maimon-Rokach Editors. Springer.
- Giudici&Figini. Applied Data mining for business and industry. Cap 5
- Albert, J. Bayesian Computation with R. (2009). Springer.

Comparación y evaluación de modelos

Tenemos dos objetivos:

- Seleccionar un modelo entre un conjunto de modelos candidatos.

- Evaluar el modelo elegido, por ejemplo estimando su error de predicción o alguna medida global.

Evaluando el error de predicción

- Error test o de generalización

$$Err = E\left[L\left(Y, \hat{f}\left(X\right)\right)\right]$$

- Error de entrenamiento

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{n} L\left(y_i, \hat{f}\left(x_i\right)\right)$$

No es un buen estimador de Err

Hay que definir la función de pérdida (L) a utilizar! Las usuales: pérdida cuadrática, pérdida de deviancia, etc.

Error de entrenamiento

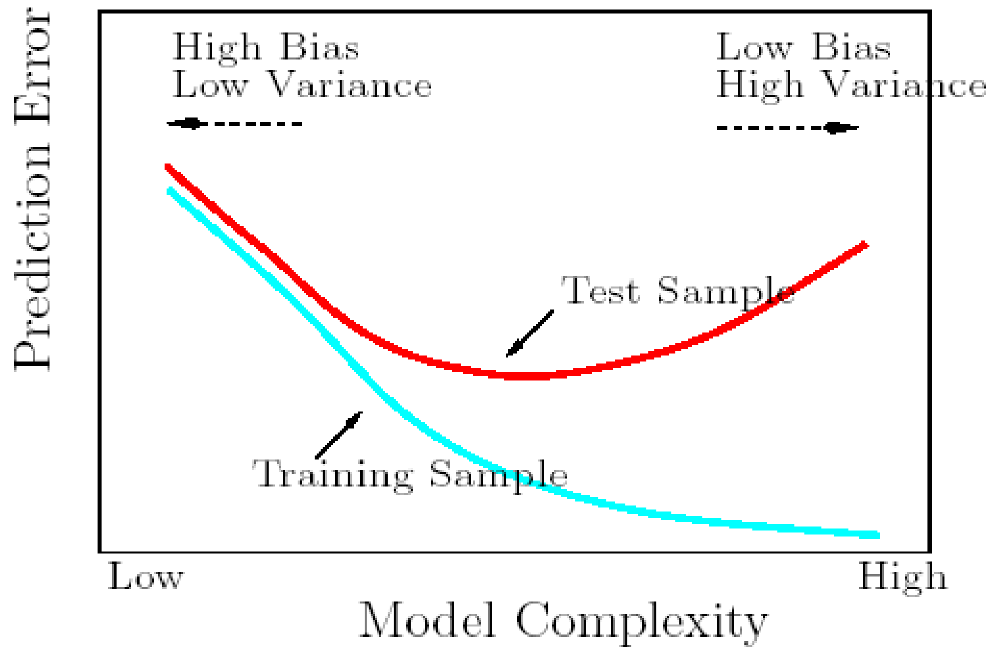


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

Overfitting!!

err decrece con la complejidad del modelo. (Fig. de H-T-F)

Funciones de pérdida típicas

Para respuesta Y cuantitativa

$$L\left(Y,\hat{f}(X)\right)=\begin{cases}\left(Y-\hat{f}(X)\right)^2 & \text{pérdida cuadrática}\\ \left|Y-\hat{f}(X)\right| & \text{pérdida absoluta}\end{cases}$$

Para respuesta G categórica

$$L\left(G,\hat{G}(X)\right)=I\left(G\neq\hat{G}(X)\right) \quad 0-1\,loss$$

$$L\left(G,\hat{p}(X)\right)=-2\sum_{k=1}^{K}I(G=k)\log\hat{p}_k(X)=-2\log\hat{p}_G(X) \quad \log-likelihood$$

Selección y evaluación de modelos

Algunos métodos de selección de modelos permiten estimar o controlar el error de predicción del modelo elegido:

- analíticamente (AIC y BIC) ó
- por re-uso eficiente de los datos (Cross validation, bootstrap)

Comparación de modelos

¿Cómo seleccionamos el modelo final? Debemos tener criterios para compararlos.

1. Basados en tests estadísticos (pruebas F, LRT, deviances o discrepancias)
2. Basados en *rankings* (AIC, BIC, etc)
3. Basados en reuso de la muestra o criterios computacionales (CV, bootstrap, bagging, etc)
4. Basados en criterios de ganancia (accuracy, ROC, Lift)
5. Basados en criterios Bayesianos (Factor Bayes)

1/2. Usando pruebas estadísticas o indicadores para comparar modelos

- Test de cociente de verosimilitud- LRT o Deviancias (para modelos anidados)
- Coeficiente de Determinación R^2 y R^2 ajustado
- Pseudos R^2 en regresión logística
- RMSE = raíz de MSE
- PRESS (Suma de cuadrados de Predicción)
- AIC, AICc, etc
- BIC
- C_p de Mallows en regresión

PRESS (Suma de cuadrados de Predicción)

$$e_{(i)} = y_i - y_{(i)} = \frac{e_i}{1-h_{ii}}$$

La medida PRESS para el modelo de regresión que contiene *p* parámetros se define por:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2$$ o equivalentemente $$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

Se elige el modelo que tiene el valor de PRESS más bajo.

Akaike's Information Criterion

Akaike (1974) define un "***criterio de información***" que relaciona la discrepancia K-L y LRT, penalizando la deviancia del modelo por la cantidad de parámetros:

$$AIC = -2\log(L(\theta \mid y)) + 2p$$

Deviancia

Da un compromiso entre ajuste del modelo y complejidad.

mínimo AIC = mínima discrepancia Kullblack-Leibler = máxima entropía

AIC: Akaike Information Criterion

Cuando la función de pérdida es la Deviance,

$$\text{AIC} = -2\log L(\theta|y) + 2p$$

(p = nº de parámetros incluído intercepto).

En particular para el caso de regresión se transforma en:

$$\text{AIC} = n\log[\text{SSEp}/n] + 2p$$

Para comparar 2 modelos, comparar AIC de model 1 vs AIC de model 2.

→ Modelos no necesitan ser anidados

→ AIC tiende a favorecer modelos más complicados

→ **MENOR ES MEJOR**

AIC y muestras pequeñas

Si n no es muy grande respecto al número de parámetros estimados, se recomienda usar AIC_c

$$AIC_c = -2\log(L(\theta \mid y)) + 2p\left(\frac{n}{n-p-1}\right)$$

En general, esto se usa si ***n/p*** es pequeño (menos de 40).

BIC: Bayesian Information Criterion o de Schwarz

$$BIC = -2\log(L(\theta \mid y)) + 2p\log(n)$$

(p = nº de parámetros, incluído intercepto).

- Los modelos no necesitan ser anidados
- Menor es mejor, como AIC
- $BIC_1 - BIC_2 \approx -2 \log(\text{Bayes Factor}_{12})$ para model 1 vs. model 2.

Observación

Los criterios AIC y Cp de Mallows tienden a dar modelos óptimos ‘más grandes’ que el criterio BIC.

Criterio de Mallows (Cp de Mallows) en regresión

Se trata de encontrar un modelo donde *el sesgo* y *la varianza* de los valores ajustados sean moderados.

Para esto, se define el estadístico $\mathbf{C_p}$ de modo de minimizar el error cuadrático medio de un valor ajustado.

Sean:

$\mathbf{SSE_p}$: suma de cuadrados del error del modelo que contiene ***p*** parámetros, incluyendo el intercepto,

$\mathbf{s^2}$: **varianza estimada** con el modelo completo.

Si el modelo con p parámetros es adecuado, $E(SSE_p)=(n-p)\sigma^2$

Criterio de Mallows (Cp de Mallows)

Se define el estadístico de *Mallows* como

$$C_p = \frac{SSE_p}{s^2} + (2p - n)$$

Ojo! Acá p = # parámetros

Se prueba que **E($\mathbf{C_p}$)=p** si el sesgo = 0.

Esto dice elegir un modelo con **p** parámetros tal que **$\mathbf{C_p}$** sea lo más parecido posible a **p.**

Datos supervisor (Chaterjee)

Table 11.4 Values of C_p Statistic (All Possible Equations)

Variables	C_p	Variables	C_p	Variables	C_p	Variables	C_p
1	1.41	1 5	3.41	1 6	3.33	1 5 6	5.32
2	44.40	2 5	45.62	2 6	46.39	2 5 6	47.91
1 2	3.26	1 2 5	5.26	1 2 6	5.22	1 2 5 6	7.22
3	26.56	3 5	27.94	3 6	24.82	3 5 6	25.02
1 3	1.11	1 3 5	3.11	1 3 6	1.60	1 3 5 6	3.46
2 3	26.96	2 3 5	28.53	2 3 6	24.62	2 3 5 6	25.11
1 2 3	2.51	1 2 3 5	4.51	1 2 3 6	3.28	1 2 3 5 6	5.14
4	30.06	4 5	31.62	4 6	27.73	4 5	29.50
1 4	3.19	1 4 5	5.16	1 4 6	4.70	1 4 5 6	6.69
2 4	29.20	2 4 5	30.82	2 4 6	25.91	2 4 5 6	27.74
1 2 4	4.99	1 2 4 5	6.97	1 2 4 6	6.63	1 2 4 5 6	8.61
3 4	23.25	3 4 5	25.23	3 4 6	16.50	3 4 5 6	18.42
1 3 4	3.09	1 3 4 5	5.09	1 3 4 6	3.35	1 3 4 5 6	5.29
2 3 4	24.56	2 3 4 5	26.53	2 3 4 6	17.57	2 3 4 5 6	19.51
1 2 3 4	4.49	1 2 3 4 5	6.48	1 2 3 4 6	5.07	1 2 3 4 5 6	7
5	57.91	6	57.95	5 6	58.76		

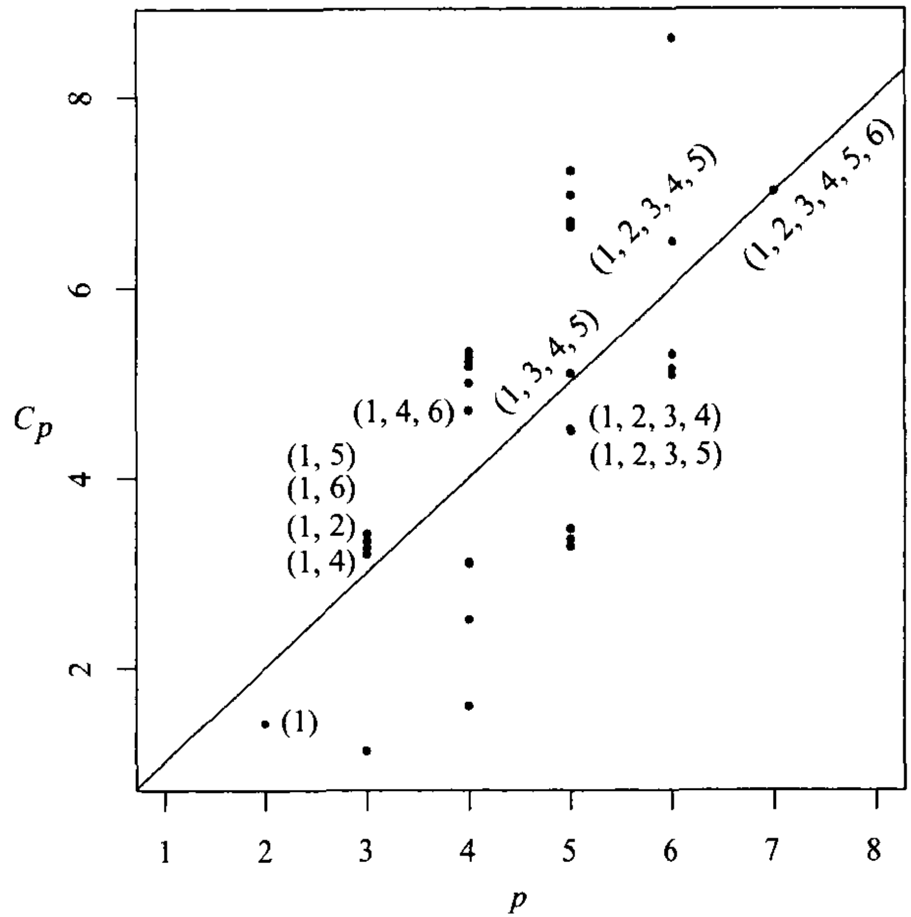


Figure 11.1 Supervisor's Performance Data: Scatter plot of C_p versus p for subsets with $C_p < 10$.

Consideraciones con el uso de Cp:

Table 11.2 Variables Selected by the Forward Selection Method

Variables in Equation	min($\|t\|$)	RMS	C_p	p	Rank	AIC	BIC
X_1	7.74	6.993	1.41	2	1	118.63	121.43
X_1X_3	1.57	6.817	1.11	3	1	118.00	122.21
$X_1X_3X_6$	1.29	6.734	1.60	4	1	118.14	123.74
$X_1X_3X_6X_2$	0.59	6.820	3.28	5	1	119.73	126.73
$X_1X_3X_6X_2X_4$	0.47	6.928	5.07	6	1	121.45	129.86
$X_1X_3X_6X_2X_4X_5$	0.26	7.068	7.00	7	–	123.36	133.17

Elegida por Forward

Elige AIC

C_p elige X1-X3-X4-X5. Distinto a lo que elige AIC o BIC. Pero Cp no es bueno aquí, no hay disponible un buen estimador de σ^2. (RMS en el modelo grande es mayor!!)

TABLE 10.1 Summary of All Possible Regressions for the Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

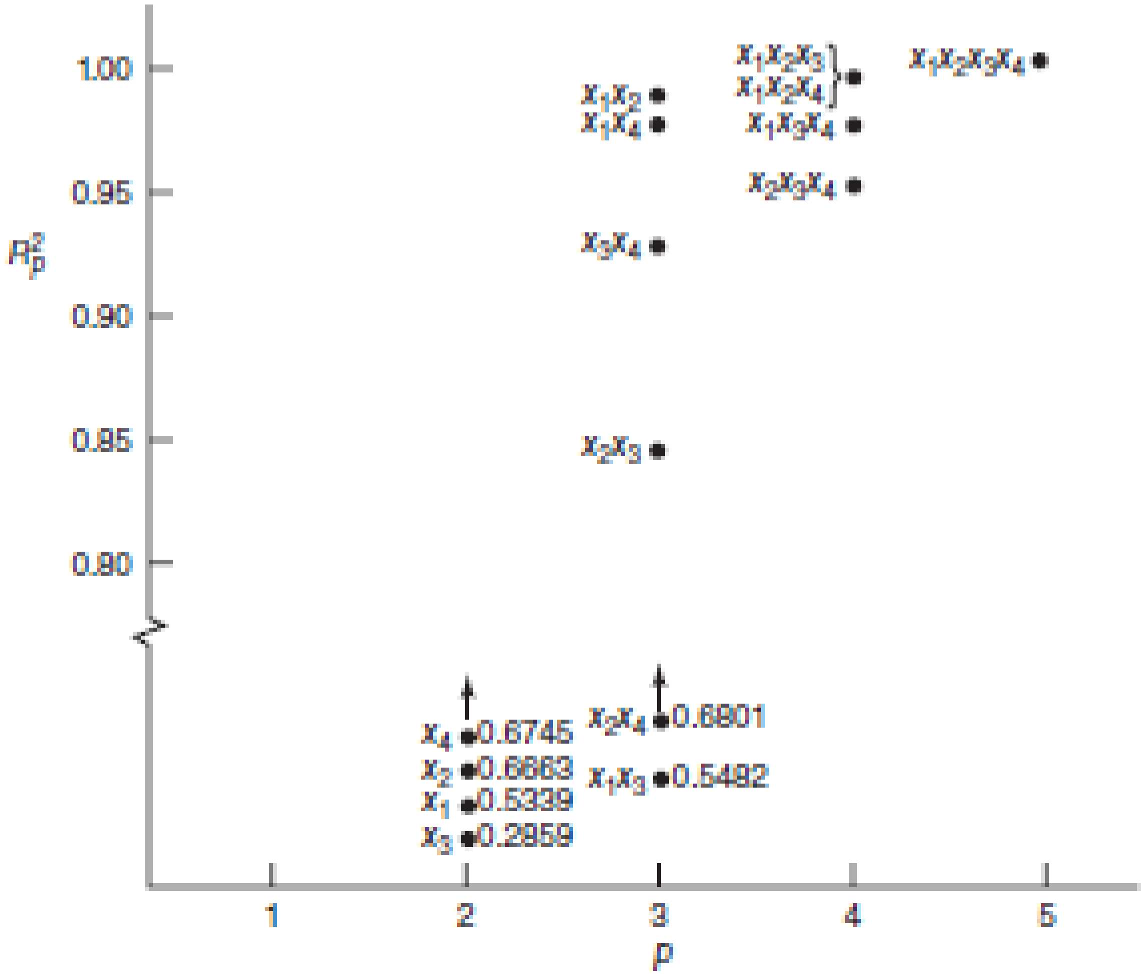


Figure 10.4 Plot of R_p^2 versus p, Example 10.1.

TABLE 10.4 Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS x_1, x_2 =	93.8827		PRESS x_1, x_2, x_4 =	85.3516

[a] $R^2_{\text{Prediction}} = 0.9654$, $\text{VIF}_1 = 1.05$, $\text{VIF}_2 = 1.06$.

[b] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

Observaciones al ejemplo de Hald

- Ambos modelos se eligieron por alto R^2
- Comparando por MSE quedaríamos con el mas chico, lo que equivale a mayor R^2aj
- Mirando Cp ambos son parecidos
- El PRESS es menor en el modelo mas grande ¿lo elijo?

OJO!! ver los FIV!! Es alto para X_4

- Entonces finalmente quedamos con el modelo más chico, con X1 y X2

Ejercicio: Autos.csv

- Evaluar modelos de regresión múltiple con todos los indicadores anteriores.
- Aplicar forward para seleccionar variables, y luego entre los modelos distintos tamaños considerar Cp, R2, Bic, .. para seleccionar uno.

3. Criterios con reutilización de la muestra: Cross Validation

El proceso de evaluar el desempeño de un modelo se conoce como “evaluación del modelo” (*model assesment*), mientras que el proceso de seleccionar el nivel apropiado de flexibilidad para un modelo se conoce como “selección del modelo” (*model selection*).

CV puede ser usada para estimar el *error tes*t asociado a un modelo, para evaluar su rendimiento, o para seleccionar el nivel adecuado de flexibilidad.

Validación cruzada

(k-CV: k-fold Cross-Validation)

Es el método más simple para estimar el error de predicción o error de generalización.

Directamente estima

$$Err = E[L(Y, f(X))]$$

Validación cruzada

- Se divide aleatoriamente el conjunto de datos en k subconjuntos de intersección vacía (más o menos del mismo tamaño). Típicamente, k=10.
- En la iteración i, se usa el subconjunto i como conjunto de prueba y los k-1 restantes como conjunto de entrenamiento.
- Como medida de evaluación del método de clasificación se toma la media aritmética de las k iteraciones realizadas.

$$CV = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f^{-k(i)}(x_i))$$

Variantes de la validación cruzada

- **“Leave one out” (LOOCV)**: Se realiza una validación cruzada con k particiones del conjunto de datos, donde k coincide con el número de casos disponibles.

(en regresión y con pérdida cuadrática, LOOCV=PRESS)

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

- **Validación cruzada estratificada**: Las particiones se realizan intentando mantener en todas ellas la misma proporción de clases que aparece en el conjunto de datos completo.

Ejemplo de ISLR con data=Auto

- Ajuste de modelos lineal, cuadrático, cúbico.

Calculo media de SSE en testing para compararlos.

- Leave one out para modelos polinómicos.

Obtener media de SSE sobre todos los ajustes.

- Idem con K-CV, para modelos polinómicos.

4. Indicadores basados en criterios de costo o ganancia

1) Calcular la matriz de confusión y evaluar con medidas que se definen a partir de esta:

a) Sensibilidad, Especificidad, Accuracy, precisión, F-measure.

b) Definir criterios basados en costo o ganancia.

	Pred +	Pred-
+	VP	FN
-	FP	VN

2) Evaluar la capacidad de predicción en algún sentido (ROC, AUC)

Métricas a partir de tabla de clasificación

$$Precisión = \frac{VP}{VP+FP} = \frac{VP}{total\ clasif\ +}$$

$$Sensibilidad = \frac{VP}{VP+FN} = \frac{VP}{total +} = Recall$$

$$Especificidad = \frac{VN}{FP+VN} = \frac{VN}{total -}$$

$$F_{\beta} = \frac{(1+\beta)^2 VP}{(1+\beta)^2 VP + \beta^2 FN + FP}$$

Media armónica entre precisión y sensibilidad

	Pred +	Pred-
+	VP	FN
-	FP	VN

Curva ROC
(Receiver Operating Characteristic)

Usado por primera vez para evaluar radares en la 2ª guerra mundial.

Se desarrolló fundamentalmente para aplicaciones de diagnóstico médico a partir de 1970 y comienza a popularizarse a finales de los 90 en minería de datos.

La curva ROC no se ata a un valor de corte para la clasificación, calculando S y E para todos los valores de corte posibles.

Como armamos una curva ROC?

eje X : tasa de falsos positivos (FP/N)

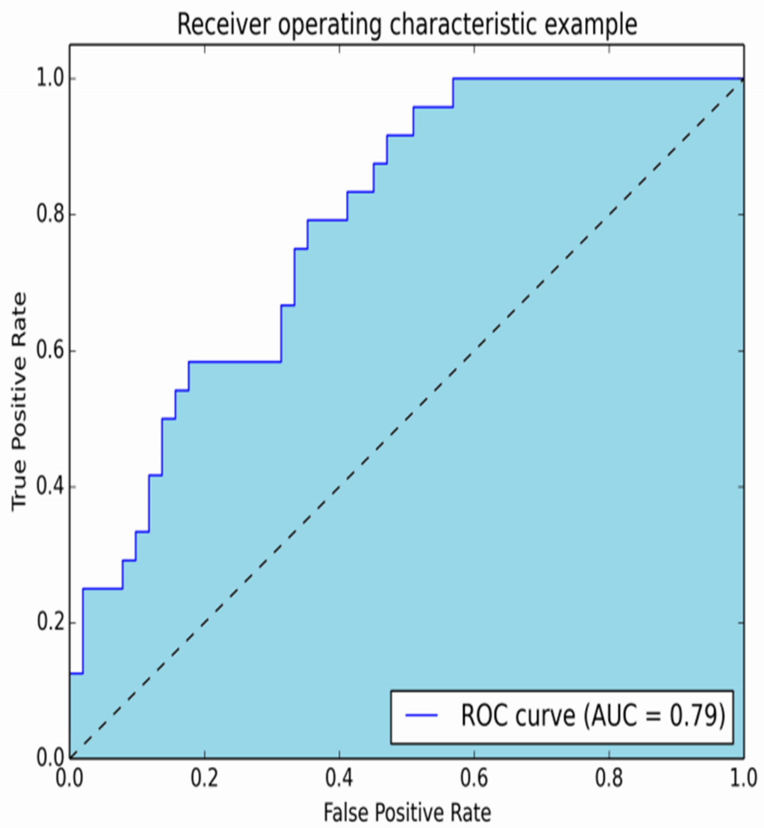
eje Y: tasa de verdaderos positivos (VP/P)

Los casos se ordenan en forma decreciente por su probabilidad de pertenencia a la clase P

La "curva" es una composición secuencial de segmentos horizontales (de izquierda a derecha y verticales de abajo a arriba)

Si el próximo caso es positivo la curva aumenta en el eje Y en una proporción de 1/P

Si el próximo caso es negativo (falso positivo) la curva se desplaza a derecha en una proporción de 1/N

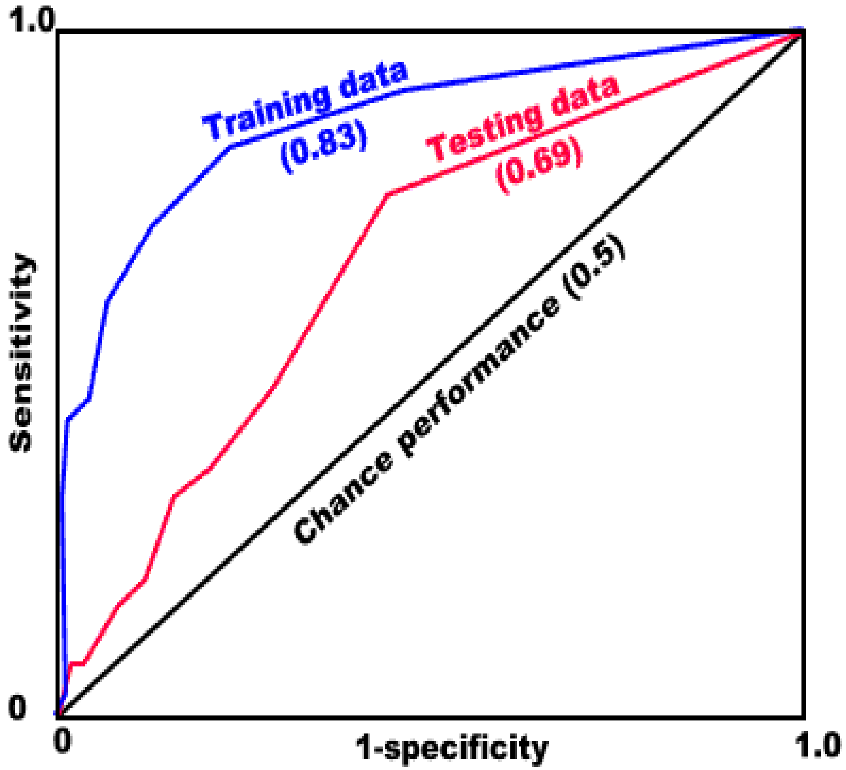


	Pred +	Pred-
+	VP	FN
-	FP	VN

Curva ROC - AUC

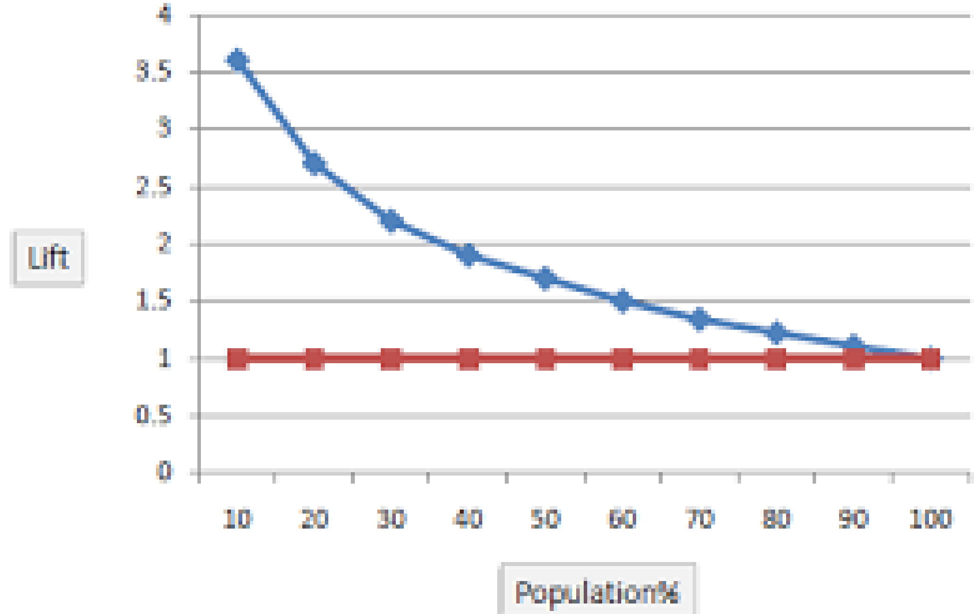
- La mayor exactitud de una prueba se traduce en un desplazamiento "hacia arriba y a la izquierda" de la curva ROC. Esto sugiere que el área bajo la curva ROC (AUC) se puede emplear como un índice de la exactitud global de la prueba.

- **AUC** refleja qué tan bueno es el test para discriminar + y -, por lo que mide **capacidad predictiva**.



Curva Lift

Lift mide la performance de un modelo predictivo, calculando el % de bien clasificados entre los primeros k casos (generalmente separados en cuantiles), ordenados de modo decreciente según la probabilidad predicha.



Métricas de comparación multiclase

La tabla de confusión sigue siendo válida, así como Accuracy.

Pero no así sensibilidad, especificidad, ROC, AUC, etc..

Se pueden hacer curvas ROC del tipo "*one vs all*" ó "*one vs one*" y evaluar AUC según estas.

Pronosticados

Reales

	A	B	C	D	E
A					
B					
C					
D					
E					

Ejemplo data=Cancer.xls

- Ajustar distintos modelos y evaluar ROC y AUC en estos.
- Graficar lift

5. Basados en criterios bayesianos: Factores Bayes.

Es el método básico para elegir modelos desde la perspectiva bayesiana. Es el análogo de los tests de cocientes de verosimilitud de la inferencia clásica.

Lo básico: la información a priori se combina con la posteriori en un cociente para dar evidencia en favor de uno u otro modelo.

Bayes Factors es un método flexible, no se requiere que los modelos estén anidados.

Es fácilmente interpretable.

Planteo general de Factores Bayes

Sean los datos X, se quiere decidir entre 2 modelos en competencia: M_1 y M_2, cada uno con parámetros θ_1 and θ_2.

$$M_1: f_1(x \mid \theta_1) \;\; y \;\; M_2: f_2(x \mid \theta_1)$$

Necesitamos distribuciones a priori para θ_1 y θ_2 y probabilidades a priori de cada modelo M_1 and M_2

Bayes Factor

El odds ratio posterior en favor de M_1 sobre M_2 es:

$$BF = \frac{P(x/M_1)}{P(x/M_2)} = \frac{\int_{\theta_1} P(x/\theta_1, M_1) P(\theta_1/M_1) d\theta_1}{\int_{\theta_2} P(x/\theta_2, M_2) P(\theta_2/M_2) d\theta_2}$$

Que también puede escribirse como

$$\text{Bayes Factor} = BF(x) = \frac{\pi(M_1 \mid x)/\, p(M_1)}{\pi(M_2 \mid x)/\, p(M_2)}$$

- Si los modelos son anidados e igualmente probables a priori, se tiene el cociente de verosimilitudes.
- Suponiendo que no hay preferencia a priori por ninguno de los dos modelos, $P(M_1) = P(M_2)$, entonces se tiene la siguiente regla para comparer modelos:

Regla para interpretar BF (Jeffreys)

$$\text{Bayes Factor} = BF(x) = \frac{\pi(M_1 \mid x)}{\pi(M_2 \mid x)}$$

Evidencia a favor del modelo M1:

Si B(x) < 1 → evidencia negativa (apoya a M2)

If 1 < B(x) <3 → evidencia escasa a favor de M1.

If 3< B(x) <10 → evidencia sustancial a favor de M1.

If 10 < B(x) <100 → evidencia fuerte o muy fuerte a favor de M1.

If B(x)>100 → evidencia decisiva a favor de M1.

Ejemplo: factores de riesgo asociados con bajo peso al nacer

Se considera la base de datos birthwt que tiene 189 observaciones y 10 variables. Proceden del Baystate Medical Center.

Se trata de relacionar la variable bwt (birth weight in grams) con el resto de variables mediante un modelo de regresión bayesiano.

Para comparar los diferentes modelos se usan factores Bayes cruzados para cada uno de los modelos considerados.

Variables:

Low= indicator of birth weight less than 2.5 kg.

age = mother's age in years.

Lwt = mother's weight in pounds at last menstrual period.

race = mother's race (1 = white, 2 = black, 3 = other).

Smoke =smoking status during pregnancy.

ptl = number of previous premature labours.

ht =history of hypertension.

ui = presence of uterine irritability.

ftv = number of physician visits during the first trimester.

bwt = birth weight in grams.

Ejemplo en R:

```
library(MCMCpack)
data(birthwt)

model1 <- MCMCregress(bwt~age+lwt+as.factor(race)+smoke+ht,
data=birthwt, b0=c(2700,0,0,-500,-500,-500,-500),
B0=c(1e-6,0.01,0.01,1.6e-5,1.6e-5,1.6e-5,1.6e-5), c0=10, d0=4500000,
marginal.likelihood="Chib95", mcmc=10000)
model2 <- MCMCregress(bwt~age+lwt+as.factor(race)+smoke,
data=birthwt, b0=c(2700,0,0,-500,-500,-500),
B0=c(1e-6,0.01,0.01,1.6e-5,1.6e-5,1.6e-5),
c0=10, d0=4500000,
marginal.likelihood="Chib95", mcmc=10000)
model3 <- MCMCregress(bwt~as.factor(race)+smoke+ht,
data=birthwt, b0=c(2700,-500,-500,-500,-500),
B0=c(1e-6,1.6e-5,1.6e-5,1.6e-5,1.6e-5), c0=10, d0=4500000,
marginal.likelihood="Chib95", mcmc=10000)
BF <- BayesFactor(model1, model2, model3)
print(BF)
```

Ejemplo

$$BF(M_1/M_2) = \frac{P(x/M_1)}{P(x/M_2)} = \frac{0.82766}{0.05878} = 14.08$$

La matriz de Factores Bayes es:

	model1	model2	model3
model1	1.000	14.08	7.289
model2	0.071	1.00	0.518
model3	0.137	1.93	1.000

Lo que indica elegir el modelo 1 por sobre los demás.

Las probabilidades posteriori de cada modelo son

model1	model2	model3
0.82766865	0.05878317	0.11354819